THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 5, May–June 1978
Printed in U.S.A.

Evaluation of a Word Recognition System Using Syntax Analysis

By S. E. LEVINSON, A. E. ROSENBERG, and J. L. FLANAGAN

(Manuscript received May 18, 1977)

A speech recognition system has been implemented which accepts reasonably natural English sentences spoken as isolated words. The major components of the system are a speaker-dependent word recognizer, a programmed grammar, and a syntax analyzer. The system permits formulation of complete sentences from a vocabulary of 127 words. The set of sentences selected for investigation is intended for use as requests in an automated travel information system. Results are presented of evaluations for speakers using their own stored reference patterns, the reference patterns of other speakers, and composite reference patterns averaged over several speakers. For speakers using their own reference patterns the median error rate for acoustic recognition of the individual words is 11.7 percent. When syntax analysis is applied to the complete sentence, word recognition errors can be corrected and the error rate reduced to 0.4 percent.

1. INTRODUCTION

A speech recognition system composed of a programmed syntax analyzer and a speaker-dependent word recognizer has been evaluated. The system accepts complete sentences in which the successive words are spoken distinctly and in isolation. The purpose of the experiment is to determine the capability of syntax analysis for improving the accuracy of word recognition and for expanding the command ensemble of a voice-actuated system.

The word recognition system, designed by Itakura,[1] is based on representing speech utterances by equally spaced frames of LPC coefficients. Recognition ensues from a comparison of a sample input pattern of LPC coefficients with an ensemble of stored reference patterns previously established by the designated speaker. The comparison consists of a frame-by-frame scan of a sample pattern against each reference pattern. A distance metric (or measure of dissimilarity) is calculated and accu-

mulated hy a dynamic programming technique as the scan proceeds. The vocahulary item corresponding to the reference pattern with the lowest accumulated distance is designated the recognized item. In addition, a distance rejection threshold is imposed. If the accumulated distance exceeds the threshold at any frame during a reference scan, that particular reference comparison is ahorted. If all reference comparisons for a sample pattern are ahorted, the result is said to he "no match" or "reject."

II. EVALUATION OF THE ACOUSTIC ANALYZER

An earlier evaluation of the automatic word recognition system was carried out over a five-month period over dialed-up telephone lines.[2] Thirteen speakers participated in that test. Each dialed the system once a day and provided utterances of words selected from an 84-word vocahulary. The 84-word vocahulary was designed to provide one-word responses to questions asked hy a computer-controlled digital voice response system. The computer was programmed to provide airline flight information requested hy a caller. In this system the question-answer dialog that takes place hetween the caller and the computer results in the specification of a category of flights for which information is desired. Because of the nature of this dialog, 50 of the 84 vocahulary items were the names of North American cities. Other entries were digits, days of the week, etc. In the evaluation using this vocahulary, with approximately 750 trials per speaker, the median word error rate was 8.4 percent. This figure is composed of 5.7 percent rejections and 2.7 percent actual mismatches.*

III. EXPERIMENTAL DESCRIPTION

The vocahulary selected for the present evaluation was designed to fulfill a similar function as that of the earlier one, namely to request flight information and to make reservations using an automated system with word-recognition capabilities. The difference is that in the present system the requests are made in the form of complete sentences rather than as one-word responses to queries. The 127-word vocahulary for this purpose is shown in Tahle I together with some sample sentence requests. The vocahulary contains many auxiliary and function-type words so that reasonahly natural English sentences may he formed. The vocahulary includes 10 city names. In the earlier mode using the question-answer dialog, depending on the complexity of the task, a long series of questions may he necessary to specify a complete request. In the present mode,

* Therefore, the term "word error rate," as used in this paper, is more appropriately defined as the rate of nonrecognition, since it includes both outright errors (mismatches) and rejects.

Table I — 127-word vocabulary for requesting flight information and reservations and two sample sentences constructed from this vocabulary

1 Evening	33 To	65 Reservation	97 Card
2 Nine	34 Charge	66 A	98 Saturday
3 October	35 Make	67 Fare	99 Pay
4 Douglas	36 Home	68 BAC	100 By
5 DC	37 Five	69 Departure	101 Ten
6 Arrival	38 Does	70 Of	102 March
7 Seattle	39 Go	71 Meal	103 Cash
8 Eleven	40 Seat	72 Flights	104 Miami
9 Los Angeles	41 From	73 What	105 Thursday
10 Friday	42 Time	74 I	106 American
11 January	43 On	75 When	107 Plane
12 AM	44 December	76 Sunday	108 Eight
13 April	45 June	77 Boston	109 Club
14 May	46 Would	78 Arrive	110 Master
15 Morning	47 Some	79 Twelve	111 Office
16 Detroit	48 Many	80 Leave	112 My
17 Do	49 In	81 August	113 Class
18 New York	50 Please	82 For	114 Six
19 At	51 Will	83 November	115 Three
20 Tuesday	52 Lockheed	84 Philadelphia	116 Washington
21 Oh	53 Want	85 February	117 Night
22 Wednesday	54 Flight	86 Are	118 Phone
23 Need	55 Four	87 There	119 Area
24 Chicago	56 Depart	88 Return	120 Two
25 September	57 Repeat	89 Coach	121 Code
26 Is	58 Take	90 O'clock	122 Nonstop
27 PM	59 Number	91 How	123 Seats
28 Boeing	60 Denver	92 Much	124 Seven
29 Information	61 Diners	93 Served	125 Times
30 Afternoon	62 Prefer	94 Credit	126 Stops
31 Express	63 July	95 The	127 First
32 Like	64 Monday	96 One	

Sample test sentences:
"I would like some information please."
"I would like one first-class seat on flight number four four to Los Angeles on Saturday the oh one January."

the efficiency of natural English is approached by combining several commands in a single sentence input. The second sample test sentence in Table I is a good example. A more complete description of the task domain and the grammar is found in a companion paper by Levinson.[3]

Seven speakers—five male, two female—participated in the evaluation. The system programs resided in a Data General Nova 840 computer. Speech was input to the system via dialed-up telephone lines from an ordinary handset adjacent to the computer console. The speakers spoke their utterances after a prompt from the console. A display scope provided an intensity curve for their current input, together with end-point markers. The speakers had the option of repeating an utterance if they felt it was botched or corrupted by external noise disturbances.

Two sessions per speaker were devoted to establishing reference patterns. In each of these sessions speakers provided a single utterance

of each of the 127 words in the vocabulary. Reference patterns were computed from these utterances. Each speaker therefore ended up with a reference containing two distinct reference patterns for each word in the vocabulary. Speakers could also provide additional optional pronunciations for the articles "a" and "the."

Finally, each speaker provided one or more test sessions in which a total of 51 specified sentences were input as strings of isolated words, for a total of 444 words. There was thus an average of 8.7 words per sentence. Every word in the vocabulary and every production rule in the grammar were represented in the sentence set at least once. The word utterances composing the sentence strings for each speaker were stored on disk files.

Recognition was carried out off-line with acoustic recognition followed by parsed recognition accomplished by the syntax analyzer. For each test sentence the acoustic recognizer provided to the parser a matrix of distances or scores $[d_{ij}]$, $i = 1,2,\ldots,N$, $j = 1,2,\ldots,M$, where i represents the ith word in a sentence string of $N \leq 22$ words, and j represents the jth vocabulary item in the vocabulary of size $M = 127$ words. The smaller the score d_{ij}, the closer is the acoustic match for the ith word in the sentence to the jth word in the vocabulary. Recognition was carried out under four different experimental conditions. Two conditions were examined in which the utilized references were those for the designated speaker. In the first of these a fixed rejection threshold was imposed. In the second there was no rejection threshold. With a rejection threshold, an arbitrarily large number was assigned to the distance score for each rejected candidate. In the third experimental condition each speaker was compared against an arbitrarily selected reference. In the fourth condition each speaker was compared against a reference which was a composite of individual references from four arbitrarily selected male speakers.

IV. RESULTS

The overall results are shown in Table II as median error rates over the seven speakers.

"Word error—acoustic best candidate" refers to the rate at which the specified test word was not the best acoustic candidate. "Word error—acoustic five best candidates" refers to the rate at which the specified test word was not included among the five best acoustic candidates. The recognition scores for word error—acoustic best candidate are quite comparable to those obtained in the earlier evaluation. Given the larger size of the vocabulary, and especially the greater frequency of common, more easily confused words, the 11.7 percent word error performance*

* Compared to 8.4 percent for an initial trial in the earlier study with an 84-word vocabulary.

Table II — Median error rates over five speakers with 51 test sentences per speaker

	Condition: Reference:	1 Designated speaker	2 Designated speaker	3 Arbitrary male speaker*	4 Composite of four male speakers*
Word error	Rejection threshold:	Fixed	None	Fixed	Fixed
	Acoustic best candidate:	11.7%	10.8%	45.5%	34.9%
	Acoustic five best candidates:	1.8%	1.1%	20.0%	20.0%
	Parsed:	0.4%	1.6%	5.6%	6.5%
	Sentence error parsed:	3.9%	5.9%	35.3%	37.2%

* Scores include female speakers using male references.

seems reasonable. As anticipated, the syntactic constraints imposed by the task language have a powerful correcting influence on acoustic word errors. For example, for condition 1 the median number of word errors was reduced from 52 to 2 out of a total of 444.† Since a single word error creates a sentence error and since the number of sentences in the sample is relatively small, the parsed sentence error rate is not as reliable an indicator of the improvement gained by parsing as the parsed word error rate. It is interesting to note that although the acoustic word error rates are about the same, with or without a rejection threshold, the parsed word and sentence error rates are somewhat larger for the no-rejection-threshold condition. We attribute this result to the following possible situation. If a specified word in a sentence string is poorly recognized acoustically, in the condition with a rejection threshold it will have the same arbitrarily large distance score as other rejected candidates. Without a rejection threshold, however, the true word may have a score which is considerably worse than other candidates resulting in a greater chance of misleading the parser.

The recognition system was not designed to be speaker-independent. We did, however, try a naive experiment in that mode. Table II also shows the results of comparing all speakers against an arbitrary reference, condition 3, as well as a comparison of all speakers against a composite reference, condition 4. It was anticipated that, although acoustic recognition would be considerably poorer for these conditions than for the speaker-dependent condition, the parser would be able to compensate to some extent this poor performance, resulting in a reasonable overall recognition performance. This seems to be true with a 5 or 6 percent parsed word error rate and 35 or 37 percent parsed sentence error rate. Comparing speakers against an arbitrary reference does not seem significantly different from comparing them against a composite reference. The performance of the two female speakers in both these condi-

† I.e., a reduction of word error rate from 11.7 percent to 0.4 percent!

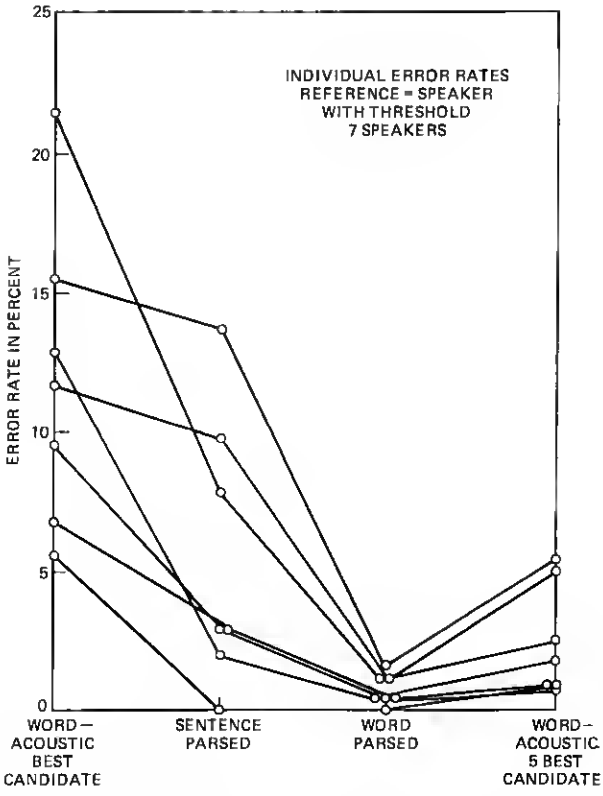

Fig. 1—Individual error rates for the speaker-dependent condition.

tions was considerably worse than that of the male speakers. The parsed word error rate for the women, for example, was approximately 30 percent.*

Individual error rates for the speaker-dependent condition are shown in Fig. 1. Most striking is the contraction of a large range of acoustic word error rates (best candidate) to a very tight range of parsed word error rates all below 2 percent. Parsed sentence error rates vary over a wide range and are sensitive functions of parsed word error rates. An additional indicator of acoustic word performance is acoustic word error rate, five best candidates. These rates occupy a considerably reduced and overall lower range than the standard acoustic word error rates. This measure may be a more reliable predictor of parsed error rates, as shown by the monotonic character of the lines that connect these individual

* The female speakers, therefore, contribute substantially to the error scores for conditions 3 and 4. This is not surprising in that the references for 3 and 4 were derived from male speakers only.

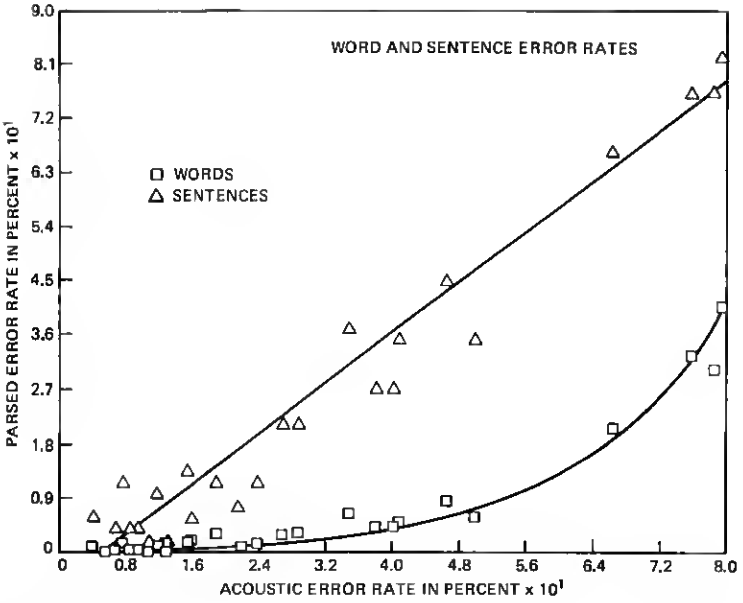

Fig. 2—Parsed error rate as a function of acoustic error rate for both words and sentences.

rates with the parsed rates. This seems reasonable since the five-candidate rate gives a good measure of the quality of the acoustic recognizer in the sense of indicating whether the true word has a good chance of having a low score.

Finally, individual parsed error rates collected from all speakers and conditions are plotted versus individual acoustic word error rates (best candidate) in Fig. 2 to characterize the effectiveness of the parser over the widest possible range of performance. This figure is analogous to the one in the companion paper by Levinson[3] which shows simulated results. The trends in both figures are the same, but the parsed error rates are significantly greater functions of acoustic error rate for the actual recognizer than for the simulation. The solid curves drawn have been fitted by eye. Note that although there is considerable scatter among the individual parsed sentence error rates the trend is almost linear. It is evident that even though the parser is a highly effective corrector of acoustic word errors this beneficial effect is neutralized to some degree by the highly sensitive dependence of sentence error on word error.

V. CONCLUSION

The command ensemble for an automatic word recognizer can be greatly expanded by forming complete sentences from a relatively modest word vocabulary. For applications where speaking discipline can

be exercised, complete sentences can be input on a word-by-word basis. The present study demonstrates that realistic syntactic constraints can dramatically compensate acoustic errors by the use of a well-constructed parser.

REFERENCES

1. F. Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition," IEEE Trans. Acoustics, Speech, and Signal Processing, *ASSP-23*, 67–72, 1975.
2. A. E. Rosenberg and F. Itakura, "Evaluation of an Automatic Word Recognition System Over Dialed-up Telephone Lines," talk presented at the 92nd meeting of the Acoustical Society of America, San Diego, November 1976.
3. S. E. Levinson, "The Effects of Syntactic Analysis on Word Recognition Accuracy," B.S.T.J., this issue, pp. 1627–1644.